

DOCUMENT RESUME

ED 248 286

UD 023 769

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE Aptitude Testing Controversy: Beliefs, Not Values, Are on Trial.
PUB DATE 84
NOTE 38p.
PUB TYPE Information Analyses (070) -- Viewpoints (120)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Aptitude Tests; *Beliefs; Cultural Differences; Predictive Validity; Scores; *Social Bias; Social Values; *Test Bias; *Test Construction; Test Reliability; Test Validity

ABSTRACT

This paper analyzes research findings about the role of values and beliefs in aptitude testing. Indications are found that the ongoing controversy over the lack of ethnic, racial, and sex fairness in aptitude testing is rooted in the tests themselves. Test designers, it is argued, are influenced by their socioeconomic position (they are usually members of the majority group), but they are also swayed by a desire for statistical elegance and scientific methodology. Recommendations are made: test makers must accommodate social differences, however difficult to quantify; and the scientific community should be more receptive and less defensive toward criticism about ability testing and show sensitivity toward the problems created by imperfect instruments. These aspects of aptitude testing receive particular attention: the significance of large group differences in scores found among different ethnic groups; the non-objective nature of many tests; the insensitivity of test makers to cultural differences; the use of test scores for job selection and admission to college; exaggerations about the usefulness of tests; reports that test limitations favor minorities; and the deliberately quantitative, empirical, and pragmatic nature of test methodology.

(KH)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED248286

APTITUDE TESTING CONTROVERSY:
Beliefs, not Values, Are on Trial

Fulogio Gonzalez-Tamayo

New York, 1984

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

E. Gonzalez-Tamayo

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as
received from the person or organization
originating it.
Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

UD023769

SUMMARY

After twenty years, the controversy about aptitude testing is not near a solution. The roots of the problem are beliefs although values have carried the argument. Conclusions about aptitudes, test scores, and performance are analyzed in terms of research findings and beliefs. Methodology has been, very often, more important than any other consideration.

INTRODUCTION

For over twenty years, we have witnessed an intense and, at times, emotional debate about the use of aptitude tests in school placement, college admission, and job selection. A wide spectrum of society has participated in the controversy: courts, politicians, civil rights and social organizations, educators, psychologists, and test developers. After much human effort and many years of polemics, attacks and counter attacks, lawsuits, and volumes of written material, we are no closer than before to decide the issues one way or another.

Like in any other controversy, emotions appeared and led both sides to support positions already taken instead of toward a search for truth. Scientists were not only annoyed but also fired up by some of the issues discussed. Attitudes, or predispositions to respond in a positive or negative way toward an issue, played a part in the discussion. Attitude is the conclusion of a syllogism with a value and a belief as premises. Some persons are opposed to abortion (attitude) out of respect for human life (value), and they consider abortion as an act against human life (belief). In a controversy where attitudes are involved, there is a tendency to argue about values even if beliefs are the point of contention. Two persons may have different attitudes (they are for or against abortion) because of different beliefs (they do or do not believe that abortion is against human life) even though both may have the same value (both respect human life).

Value is a cognitive category with some kind of goodness attached to it (freedom of speech, democracy, merit system are good and desirable goals). Values are deeply rooted in human nature, are difficult to change, and are generally accepted by a community as a whole. Because of their general acceptance, values make an argument stronger; because of their aspect of goodness, they make the opposition unattractive and in a heated discussion there is a natural inclination to emphasize them.

Beliefs represent the union between two concepts but the union is not one

of cause and effect (lack of breathing produces death) nor has it any demonstrated empirical support (punishment is not the opposite of reward). Beliefs represent personal experiences, personal needs, or personal trust in an authority. Because of these unscientific characteristics, most people do not resort to beliefs in an argument.

In the controversy about aptitude testing, basic beliefs have been challenged by raising questions about merit system of selecting people, about the objectivity and unbiasedness of tests, about the fairness of the decisions based on tests. There are indications that the challenges have been perceived as attacks against the basic values of society; Linn(1982) and Weitzman (1982) talk about attacks against the merit system, Lerner (1979) about attacks on the truth, and a good number of others (Tenopyr, 1981; Goldman & Widauski, 1976; Novick & Ellis, 1977; Petersen & Novick, 1976; Schmidt & Hunter, 1981) labeled the attacks as sponsoring quotas, preferential treatment, or reverse discrimination. The reality is that beliefs, not values, have been the target.

Humanity has been able to make progress because beliefs were questioned at one time or another and new laws were discovered. Every scientist is well aware that progress in his own field was only accomplished after some deeply-rooted preconceptions, assumptions, and beliefs were challenged and the way was set for new research. The controversy about aptitude testing could be the opportunity for newer and sounder avenues in the search for truth.

TEST SCORES AND APTITUDES

The attacks against the test industry were started by civil rights, social and political organizations because of the adverse effects caused by tests in certain ethnic groups. This fact apparently confronted test developers with contradictory demands coming from scientific and social considerations and may have given the impression that the arguments against aptitude testing were based more on social and political motives than on scientific reasons. For Wigdor and Garner (1982) the controversy involved "polite disagreements among professionals about abstruse technical questions and heated public debates about the social implications of testing (p.7)". For Jensen (1980) the opposition to aptitude tests is mainly based on bad experiences caused by tests; "people understandably tend to move out of situations highlighting comparisons unfavorable to their self-esteem (p.21)"; by the same token, we can also say that it is understandable that people tend to stick to situations highlighting comparisons favorable to their self-esteem.

Few, if any, social changes come by accident, suddenly, or happen in a vacuum without any social or political pressure. We react defensively to any suggested change in our way of thinking, behaving, and believing; we use all our power to stop, postpone, or frustrate any change. Power is a necessary but insufficient condition for success in the long run. The attacks against the aptitude tests are still alive because there are serious and reasonable technical questions about the claims made by and the use of aptitude tests.

At the core of the controversy is the explanation of large group differences in aptitude test scores found among different ethnic groups. It is a historic fact that the majority has constantly scored higher in aptitude tests than any other minority. One wonders if the reaction from test defenders would have been the same if minority groups had scored higher than the majority in some aptitude tests. Facts are facts and we can not argue with them. What we do with the information, the kind of questions we ask, the inferences we make will depend, at

least in part, on our beliefs and experiences; either we believe in equality among different groups, without sponsoring identity or the egalitarian hypothesis (Jensen, 1980), and question the test results or we believe that the differences are real and no further investigation is needed.

A quite often heard reaction to this finding has been that differences in a test do not indicate bias or unfairness by themselves or that we can not rule out the possibility of real differences among groups. The reasoning is correct but it is also true that differences in a test do not indicate by themselves differences in aptitude. This is the basic issue of the controversy, and the available empirical and theoretical evidence cast some doubts about the trustworthiness of the information provided by a test when different groups are compared.

Tests as Objective Measures

An early belief of test developers was that intelligence and aptitude tests were objective measures of abilities independent of one's culture. We can apply to culturally different groups the same principles that Jensen (1980, p.621) used to defend sex corrections in the tests: (1) test items are not independent of sex differences, (2) tests are made up by men, (3) test bias can obscure a true difference or produce an artifactual difference (4) there is also the problem of obtaining equally random or representative samples of males and females.

The history of culture free tests has been a futile attempt (Samuda, 1975). After 25 years of efforts to develop a culture free test, Goodenough and Harris (1950) concluded that tests are appropriate for ranking individuals within a highly homogeneous group and not for ranking groups.

Aptitude tests pretend to measure steady personal characteristics or traits that are used in future performance. The traits are constructs, theoretical and hidden units, that we approach indirectly through events which are manifestations of the trait. This is the state of mind when we start developing a test and at the end of our effort we are unable to claim that the test is measuring the trait

we had in mind (Hogan, DeSoto, & Solano, 1977). Between the initial idea and the final product there are so many assumptions, limitations, obstacles, informed but subjective decisions, that the only claim we can make is that whatever the test is measuring, it will be helpful in making some predictions about other scores.

The trait has a space or domain where it is manifested. How large, varied, or uniform the domain is nobody knows. The broader the construct, and all aptitude tests include broad concepts, the more difficult it is to put limits to the domain or to define the critical features of it and, therefore, it is also more difficult to include the trait in the test. As a result, we often take the risk of giving to a very limited set of events all the meaning associated with the concept (Messick, 1980) interjecting our own beliefs in our conclusions.

The items have to be observable before they are measurable and the concept can be reduced to trivial things because many of its manifestations are not observable. Another limitation is that we are constrained by our experience when looking for manifestations, incidents, or events where the trait is shown; it is more likely that we will be able to include events we have seen or experienced than events we never experienced. The test constructor can go beyond his experience but it will be harder and less often than going to the events with which he is familiar.

There are plenty of suggestions in the literature about the relationship between items in a test and culture. For Berry (1969), events (items) can be interpreted and make sense only within the environment they were defined; for Irvine (1969), it is more valid to see abilities as personal traits intimately related to than independent from a cultural context; for Goldschmidt (1966), any comparison between groups is basically false because we compare things that are essentially not comparable; for Holtzman (1968), the more complex the behavior we are trying to measure the more important the cultural variables are in determining the behavior; Israel & Rommeveit (1954) used the phrase "divergent situational factors" in the sense that two persons may start early in life with the same degree of a particular ability but the different environmental situations in

which they live lead them to develop the abilities in different areas and in different ways. Incidents are not culture free but rather culture similar, and the culture of the test developer may have some impact on the items selected, the importance given to a part of the domain, the way in which an item is expressed.

From all the available incidents, only few can be selected and included in the test for the try-out. The selection is a subjective decision based on the resourcefulness and culture of the item writer. Data come to us only as an answer to a question, they represent information about the things we want to know; and the question, expanding what Kaplan (1964) said some years ago, reflects our values, beliefs, experiences. To state, as the National Commission on Ability Tests (Wigdor & Garner (1982, p.25) does that an ability test "yields a sample of what the individual knows and has learned to do" is misleading. Sample is a technical word with the connotation of scientific selection; a test is merely a part of a particular domain or experience; different test developers come out with different items (parts) of the same domain or ability and no test developer claims any scientific sampling of the items included in a test.

Group Differences in Test Scores and in Ability

In comparing groups in whichever ability the test is measuring we need to be confident that differences in test scores represent differences in ability. This is crucial when some irreversible decisions are made affecting the future of a person as it is the case in job selection and admissions to school.

Group Differences in Test Scores are Too Large

Most studies have found that the mean difference between majority and minority groups (generally blacks) is about one standard deviation; roughly, about 50% of the minority group score as low or lower than 16% of the majority group. We can look to these facts and say that as long as ethnic minorities have a large proportion of less educated and disadvantaged members than the general population,

those social facts are going to be reflected in test scores (Jensen, 1980; Wigdor & Garner, 1982) and avoid the issue or we can take a closer look at such differences and study their reasonableness. The differences may be real but before accepting them as such, it is necessary to have strong evidence, stronger than the evidence provided by tests. The real issue is not whether two different groups are equal in ability, they probably are not, but rather whether or not the differences found in a test are larger than the actual differences in ability among different groups. Given the information about the available pool of people taking a test, the found differences seem unreasonable.

There is a substantially greater minority than majority shrinkage in the available pool of people for every step of the educational ladder. Smaller numbers of minority than majority students go to high school, smaller numbers finish high school, smaller numbers go to college, and smaller numbers graduate from college. The statistics of drop-outs demonstrate this fact without going into the causes. The more people drop out in one stage, the fewer the number of persons available for the next. If we assume, as it seems reasonable, that the great majority of drop-outs are those falling in the lower part of the aptitude scale (low scorers), it is obvious: (1) the average score for the majority at each stage of the educational process should not change drastically because most of them go to the next, (2) substantial increases are expected in the mean score for the minority group the further they go into the educational process, because most of the low scorers are eliminated before finishing the previous stage, and (3) it is also expected, as a result of the two previous considerations, that the mean difference between the groups will be reduced the further they go in the educational process. Aptitude tests have not shown a progressive reduction in the mean group difference; studies have reported very similar group differences at different stages of the educational process. It may be the case that the differences are the same throughout the whole process but before accepting this as a fact, there is a need for better information than the one provided by the tests.

There is a similar but reverse situation in employment. Tests are given

almost exclusively for low entry level jobs (Printice Hall Report, 1975; Wigdor & Garner, 1982). Most jobs beyond the entry level require some college education, and most of the jobs go to the majority by the mere fact of numbers: many more of them than of the minority finish college. This means that a smaller proportion from the majority than from the minority are available for these jobs; additionally and very likely, the persons available from the majority are the least educated and the lower scorers in the aptitude scale, but most of the minority are still available. It is difficult to accept without strong and reliable information that the majority still scores higher than the minority and with the differences between the two groups being very similar to the differences found before the reduction in the available pool of people.

There are serious and reasonable questions about the instruments showing this constant pattern of majority intellectual superiority unless we believe that every member of the majority is superior to every member from the minority.

Sex and Ethnic Differences

Test developers, to complicate things more, made corrections for possible sex bias when differences between sexes were found. The corrections were made on the assumption or belief (Jensen, 1980, p.623) that the sexes do not really differ in general intelligence. The Stanford-Binet and the Wechsler, the two most widely used tests of intelligence, have explicitly tried to minimize sex differences in total score by discarding those items that showed the largest sex differences in difficulty (simple number of persons passing an item). They did the corrections without any sophisticated transformation of difficulty in search for equal units, or to account for different levels of ability, or in quest for invariant statistics. These corrections were made many years ago (McNemar, 1942) and no change has been made to accommodate the practice to the new techniques of item bias analysis and all the technical questions raised during the controversy involving different ethnic groups.

The differences in test scores between different ethnic groups are much larger

than between the sexes and larger than one can expect given the educational and social background of various ethnic groups. The response from the scientific community by and large has been more research or try to find sophisticated psychometric techniques that would indicate bias in the items. As expected, no new discoveries have been made, no technique has been found satisfactory to everybody, confusion exists about the interpretation of biased items, and we continue using the same tests until better techniques are discovered or until the result are more interpretable. Some of the reason for this differential response to sex and ethnic differences is beliefs in equality between sexes and inequality among different ethnic groups and that the found differences are real. It is not an accident that all the techniques about bias and fairness models appeared after the passage of the Civil Rights Act of 1964.

Insensitivity to Cultural Differences

Sherwood & Nataupsky (1968) found that conclusions about differences in intelligence between whites and blacks could be predicted from the researcher's biographical data; the correlation between the conclusions and the personal data (.45) is as good or better than the validity of the tests used in admissions, placements, and job selection. Many conclusions based on test scores are wishful thinking, a confirmation of the researcher's beliefs rather than empirical or solid evidence.

Professional who are familiar with the bilingual tests used in the United States are also aware of their technical shortcomings (Olmedo, 1981) and some of them are a disgrace to psychology. One gets the impression that anything developed in English or by American professionals is equally appropriate for any other culture and, therefore, no extra or additional work is needed. There is certain arrogance in some test developers, as in some other manifestations of American society, toward the needs and demands of other cultures. Gonzalez-Tamayo (1980) labeled two bilingual tests widely used in New York City (The Test of Basic Concepts,

Boehm, 1971; Language Assessment Battery (LAB), Board of Education, NYC, 1976) a joke and an insult to the Hispanic community. In the TBC, the translation is not very professional with unacceptable gross mistakes; there are questions about the similarity of the trait in both languages and, therefore, the usefulness of the test; the equivalence of some of the items is also questionable or clearly wrong. In the LAB, most of the items are silly and can be answered without knowing the language; it is unclear what some of the items are trying to measure; some of the alternatives do not make sense; many wrong Spanish words are used; it seems that test developers wanted to measure the problems they encountered when learning Spanish rather than the language ability of a native Spanish speaker.

APTITUDE TEST SCORES AND PERFORMANCE

The strongest criticism of test scores was directed toward their use for job selection and admission to colleges. The only legitimate reason for the use of tests in this respect is their predictive power of anticipating future performance as reflected by the validity coefficient and the regression line. The business of forecasting is not very accurate or reliable especially when dealing with human behavior.

Aptitude and Performance

The purpose of selection is to identify excellence in performance before the candidates are hired or admitted to school; this is a value that any fair minded person will accept. The points of controversy are either we believe that successful performance is related to skills and some other personal characteristics not measured by tests or we believe that tests can identify the best prepared to perform well in a given role (Wigdor & Garner, 1982, p.23) and that motivation and industry are greatly overrated, as compared with capacity, by most persons (Jense, 1980, p.22).

Success in life is not likely to be highly correlated with academic success or performance on an aptitude test (Green, 1978; Linn, 1982). Test scores predict grades in school or in a training program; the games people are required to play in aptitude tests are similar to the games teachers require in the classroom (McClelland, 1973). The link between test and non-test behavior may be more tenuous than has been generally assumed (Cole & Al., 1978). Miniature job training has superior predictive validity than aptitude tests (Siegel, 1983) for those considered by a paper and pencil test as not eligible for training.

Tests indicate general tendencies rather than rigid assessments of future performance. The longer the span between tests and performance, the lower the

the relationship and greater the importance of other factors. Most validity studies measure performance after training, after a few months on the job, or after the first year of college. Homework turned out to be the single most powerful school variable in the entire array of such variables, not only as a predictor of achievement but in virtually every other subject studied (Wolf, 1977).

Adequate Prediction from Tests

Aptitude tests, in addition to the limitations mentioned before of short span of time between test and performance, show a very modest correlation coefficient and the criterion of performance is very often supervisor's ratings or scores on a training program rather than actual performance on the job. In addition, the correlation coefficient is very unstable both in job selection and college admissions.

In employment, it has been reported ranges for the validity coefficient from the 20s to the 50s (Schmidt & Al., 1980, 1981a, 1981b, 1981c; Gael & Associates, 1972, 1975a, 1975b) with most coefficients being in the 20s and 30s. Giselli (1966) reported that, for all occupations, the average validity of employment tests for predicting success in a training program is .30 and for predicting proficiency on the job the coefficient drops to .19; in a later report (Giselli, 1973) and for a small number of jobs, he reported validities of .45 for training and .35 for job performance. The discouraging common finding has been that aptitudes tests do not explain very much (only about 9%) of the variation of performance as measured by a criterion of convenience.

In admissions to college the situation is not more encouraging. Goldman & Slaughter (1976) reported coefficients from the 10s to the 60s depending on major of study. Goldman & Widawski (1976) concluded that the additional use of the Scholastic Aptitude Test (SAT) produced virtually no increase in valid prediction over the high school point average for either blacks or chicanos, with the reported correlations being from .20 to .41. Munday (1970) reported validities

of .29 to .80 in college depending on range of talent, ability level, size of freshman class, and whether or not the students lived under college supervision. Kaplan (1982) reported an increase in percentage of students who will succeed in college of only one or two percentage points between high school grades and the addition of the SAT scores.

No person would put too much trust in a professional making decisions about his future based on so weak and unstable information. The general belief of test defenders is summarized in the words of the Commission on Ability Tests (Wigdor & Garner, 1982); "When the validity has been established that it measures adequately what it purports to measure, admission tests predict grade performance reasonably well for the first year but decreasingly well for successive years (p.13)". Validity for a test is never established (Cronback, 1971), what is adequate is a question of personal preference, reasonably well is the largest prediction we have been able to make but not necessarily a good prediction, and decreasingly well (but still well) for successive years is a statement that does not respond to the available empirical data.

EXAGGERATIONS ABOUT THE USEFULNESS OF TESTS

Tests provide some useful information, but limited, and we have to live in some kind of uncertainty always painful but especially when institutions and individuals expect from tests more than they can deliver. Since we can not measure all the important things that we and others need, want, or expect, we tend to consider what we can measure as all important (Flaugher, 1978).

Generalizability of Validity

Recent work by Schmidt, Hunter, and Caplan (1981), Schmidt, Hunter, and Pearlman (1981), Schmidt, Gas-Rosember, and Hunter (1980) in the area of generalization of the tests used in employment has been heralded by the authors and others (Tenopyr, 1981) as a break through in the area of employment testing and it got a sympathetic ear from the National Commission (Wigdor & Garner, 1982) although with reservations about some of the claims made by the researchers. The discoveries can also be seen as an indication of how modest the efforts in selection have been in producing reasonable and meaningful results. The whole thrust of the theory and research on generalizability is that aptitude tests used in employment can predict performance in different jobs and also in different tasks within the same job category.

If the selection is going to be based exclusively on test scores, as the authors sponsor, the best scorers should be placed on the first available jobs even if the applicants do not have the special skills required by the job (aptitude tests measure only general ability); a secretary could be placed as a programmer, or a bad programmer ahead of a good one if they score higher on the aptitude test.

The findings indicate how little instead of how much is known about performance. Evidently, it is related to aptitude in some degree but in a very limited

way. The little thing we are able to measure is the same in every job and probably there are many other characteristics that are also shown in every performance. These common characteristics are in addition to not instead of other special skills.

Test Scores and Increase in Productivity

Ranking of persons by aptitude scores is done in the civil service system and it is also suggested in employment (Schmidt & Al., 1979; Schmidt & Hunter, 1981) and even a formula is proposed (Schmidt & Al., 1979) about the increase in productivity that would occur if the best performers on ability tests were hired. Regardless of how sophisticated the explanations may be, the naked truth is that beliefs rather than hard data are behind these conclusions. According to Wigdor & Garner (1982, p.10) "it is believed that decisions based on this kind of selection (from test scores) will contribute to the overall performance of the institution or employer" (underline and parenthesis added).

The mistrust about the criterion and the high school point average and the complaints about their unreliability are based on the fact that, very often, they are subjective measures as opposed to the objectivity provided by a test. In developing the standard deviation for the formula of increase in productivity, the authors used several subjective measures: the judgement about good performance and the estimate of the amount of money saved by good performance. There are also the assumptions that changes in productivity are due exclusively to employee's performance and that good performers on the job are those scoring higher on the test which is not obvious and no proof is given. With all these questionable assumptions and subjective interpretations, the statement about the billions of dollars saved by hiring applicants from top to bottom are merely personal beliefs.

Test Scores and Decline in Productivity

Two statements by Schmidt and Hunter (1981), accepted by Tenopir (1981), are outrageous besides being gratuitous and unrelated to the evidence presented. The statements, again, represent beliefs rather than facts.

The decline in productivity has as a possible cause, according to the authors, the decline in the accuracy by which people are sorted into jobs due to the decline in testing. Literally, the statement is correct because almost anything is possible or few things that we know of are not possible; the authors probably have in mind more than just making a meaningless statement. The difference between scientific and common knowledge is the difference between possibility and probability; possible statements are what the man on the street makes every day.

The only facts available to them were the simultaneous occurrence in time of the decline in productivity and the decline in the use of aptitude test. Any elementary book on research would reject this unscientific and simplistic explanation of the simultaneous occurrence of two events. There are limitless possible explanations for the decline in productivity based on the authors' reasoning and one of them could be the increase in the number of psychologists used by the companies indicating how harmful psychologists could be to the industry. Thirty years ago when the productivity was high tests were not as widely used as they are today, and it clearly indicates that the use of tests and not their decline is the possible cause for the decline in productivity. Tests are mainly used for entry level jobs and it is a little simplistic to blame these employees for the decline in productivity without any higher ups.

The GE company found that many people hired under the EEO Guidelines were not promotable, is another statement in the same article. We do not know the number of many, we do not know if they were more after than before the Guidelines, we do not know if every body was promotable before the Guidelines. Besides, many

more things happened in the country and in the company during these years than just the abandonment of tests and all these things may have an impact in the promotability of the employees.

Blaming a weak person or group for the social evils, with no reliable information to support the charges except rationalizations, is a fact in the history of racism. By a previous article (Hunter & Schmidt, 1976) we know that the standards were lowered in order to hire more blacks. The decline in productivity is a complex industrial, commercial, economic, and social problem that is not going to be solved by the use of tests for entry level jobs.

TEST LIMITATIONS FAVOR MINORITIES

Test limitations, like unreliability, guessing, bias, and invalidity have been known since the inception of testing. These limitations may impact differently on different groups and this is the case but researchers report that the limitations favor the low scoring groups and the issue, therefore, has not been a determining factor during the controversy. The problem is that the conclusions were based on reasonable but theoretical assumptions and the reality may be different.

Guessing

Guessing is present in most aptitude test and its amount depends on the number of alternatives presented to each question, the difficulty level of the item, the ability level of the group and some other not easily identifiable factors.

It was originally believed that guessing was related to ability and that if it was higher among high than among low scoring groups it would indicate differences in ability rather than just chance differences or test wiseness. There is evidence that this is not the case. Carlson (1969) showed that the two most important factors in guessing were age and socioeconomic level with intelligence as a distant third; Vaudenberg (1967) found that the first factor in mental ability was language of origin, the second verbal ability, and mental ability came only as a third factor.

It was also believed that multiple-choice questions penalized the brighter students (Wigdor & Garner, 1982; Hoffman, 1962) who tend to see more possible associations and are, therefore, attracted to misleading alternatives. Empirical data for the SAT (Levine & Drasgow, 1983) have shown the opposite: bright students select among only one or two alternatives while low scoring students select among all the choices; guessing is, therefore, much higher among high than among low scoring groups.

Unreliability and Invalidity

An enormous number of studies have been conducted with the regression line and the findings have been very consistent across groups, across tests, and across uses of the tests. The two main findings have been, (1) research has consistently been unable to prove any systematic under- or over-prediction in either the minority or the majority groups and (2) the common regression line results in overprediction of average minority criterion score.

These are facts and no fair minded person can argue with facts. Some conclusions drawn from these findings are, however, arguable to say the least. Based on the finding that the major group regression line over-predicts the minor group criterion scores, the same happens with the common regression line, some researchers (Cleary, 1968; Petersen and Novick, 1976; Jensen, 1980; Anastasi, 1982) have concluded that the common regression line favors minorities. Jensen (1980) went further and asserted that unreliability and invalidity in a test favor minorities and the elimination of these shortcomings in a test will result in the selection of fewer minority persons.

The regression line, like any other statistic in the classical test theory, is group dependent. When the mean and the standard deviation on the predictor scores are equal to the mean and standard deviation on the criterion, the regression equation is given by

$$Y = r.X + (1 - r)\bar{X}$$

which is the same formula used to predict true from observed scores, where r would be the reliability, Y the true score, and X the observed score. This equation is intriguing and it tells us that the prediction is related to the observed score and the group mean in a positive way; the higher the group mean, the higher the predicted score. The same person will have a higher predicted score in the high than in the low scoring group. This fact makes some people uneasy with the

regression line when comparing two different groups because group membership is an important factor in the prediction and it is more important the lower the validity or reliability is (r).

The unreliability of the instrument, we could say the same about the invalidity, introduces errors in the prediction and the errors are positively related to the observed scores. The estimated amount of error for a particular score is given by

$$e = (1-r)(X - \bar{X})$$

The formula tells us that the amount of error (1) is inversely related to the reliability and the higher the reliability the smaller the amount of error for any particular score, (2) depends on the difference between observed scores and the mean, being the errors positive when the observed scores are above the mean and negative when the observed scores are below the mean, (3) the errors, therefore, always favor (are positive) those who score above the mean and they are more favored the higher their observed scores are, (4) the errors always disfavor (are negative) those scoring below the mean and they are more disfavored the lower the observed scores are. These observations indicate that the difference, the true difference, between low and high scorers is doubly amplified by the unreliability of the instrument; first, the ability of the high scorers is inflated (positive errors) and, second, the ability of the low scorers is deflated (negative errors). The unreliability, clearly, favors the high scoring individuals contrary to Jensen's (1980) position.

Based on the previous considerations, it is possible to say that the common regression line when the reliability and validity is less than perfect and under the condition of minimizing the errors of prediction (best fit regression line) never systematically over- or under-predicts the scores of any group except in the unlikely event that all the members of a particular group score either above or below the group mean. By definition of best regression line, all the scores above the mean are underpredicted and all the scores below the mean are underpredicted. In the formula for estimating the true or criterion scores (Y) from

the observed or predictor scores (X) we can see that when the reliability is one, the observed and the true scores are the same ($Y = X$) and the graphic representation is a straight line at 45° as shown by line 1 in figure 1. When the reliability is less than one, the line becomes flatter as represented by line 2. It seems that the estimated true scores for those scoring above the mean are underpredicted and those scoring below the mean have over-predicted true scores but there is no systematic under- or overprediction of scores.

Figure 1 about here

For each observed score above the mean, a lower true score is predicted with unreliable than with reliable instruments and a higher true score is predicted for each observed score below the mean. For an observed score of X_H , the corresponding true score would be Y_H with completely reliable tests and Y_A with unreliable tests; being $Y_A < Y_H$, the unreliability seems to underpredict the true scores for those scoring above the mean. For an observed score of X_L the corresponding true score would be Y_L with completely reliable instruments and Y_B with unreliable tests; being $Y_B > Y_L$, the unreliability of the instrument seems to overpredict the true scores for those scoring below the group's mean.

The conclusion from these analyses seems to be that the unreliability favors the low scoring groups or individuals. But this is a fallacy because the problem is not with the true scores but rather with the observed scores. The observed scores contain a certain amount of error and the error is positively related to the observed scores. The scores above the mean are inflated because the error is positive and the observed scores below the mean are deflated because the error is negative. The observed scores, in addition to the true ability, include positive (are inflated) or negative (deflated) errors. An score of X_A in a completely reliable test (true score) is increased to X_H by the unreliability of the instrument; an observed score of X_B in a completely reliable instrument (true score) is decreased to X_L . A Person with a particular level of true ability,

either Y_A or Y_B , will score X_A or X_B in a completely reliable test and X_L or X_H in an unreliable test; what has changed is not the true ability but rather the observed scores.

The difference between the predicted scores of two individuals, or the predicted means between two groups, is given by

$$Y_H - Y_L = r(X_H - X_L)$$

indicating that the predicted difference is smaller than the observed difference due to the unreliability of the instrument. The low scoring individuals, or groups, are better off with predicted than with observed scores which is not the same as saying that low scorers, individuals or groups, are favored by the prediction. We can also say that the observed differences between individuals, or groups, are larger than the true differences. From the finding that the differences in observed scores is larger than the difference in predicted scores we can not conclude that the unreliability favors the low scoring individuals or groups because the change is not in the true scores, predicted, but rather in the observed scores. There is more reasonable the conclusion that the observed scores favor the high scoring individuals or groups and are against the low scorers.

The unreliability also makes that fewer low scorers be selected contrary to the conclusion made by Jensen (1980). If the difference in true scores between two individuals, or groups, is one standard deviation with the common mean being exactly in the middle point between the means of the two groups, the difference in observed scores will be one standard deviation when the reliability is perfect, it would be 1.25 SD when the reliability is .90, and 1.50 SD when the reliability is only .50. If the cut-off score for being selected were the mean of the high scoring group, about 50% of the majority and 16% of the minority would be selected if the reliability is perfect; if the reliability is only .50, the rates of selection would be 50% for the majority and less than 7% for the minority.

The unreliability does not overpredict nor does it underpredict the true scores but rather it inflates high observed scores and deflates low observed scores

and the result of unreliability is an exaggeration of the true differences between individuals and groups. Regardless of the way we look at the unreliability it always favors the high scoring individuals or groups. The overprediction and underprediction of true scores is misleading. If we want to make the largest overprediction as a way of favoring the low scoring individuals or groups, the lesson would be to make the test completely unreliable; the proposition, besides of making no sense, would be deceiving because there would be very little to be predicted because fewer and fewer individuals from the low scoring group would be selected. By the same token, a completely unreliable test, would apparently penalize the high scoring individuals and groups by making the largest underprediction, but what kind of penalty would it be when more and more of them would be selected?

PREMINENCE OF METHODOLOGY

Methology could be elevated from handmaiden to the master of the process of inquiry (Nisbert, 1976); mental measurement has been deliberately quantitative, empirical, and pragmatic. The choice between two alternatives has been often determined by statistical elegance rather than by any other consideration.

Test Scores vs. High School Grades

There is a large body of research showing that grade point average (GPA) from high school are better predictors of college performance than the scores on the Scholastic Aptitude Test (SAT). Do colleges rely more on the GPA than on the SAT scores to make their selection decisions? Current college admission practices tend toward overuse of the SAT (Weitzman, 1982); although high school grades appear to be the predominant academic predictor, the SAT is the predominant determiner of college admissions. Needless to say that the SAT has more elegant statistical characteristics than the GPA scores.

ETS (1980) reported validities of .52 and .41 for the GPA and the SAT scores respectively, indicating that the SAT scores account for less than two thirds of the variance in performance (17%) as compared to the GPA scores (27%). The SAT, in addition to being the second best predictor, does not add too much to the prediction made by the previous high school grades. Even with the most favorable figures provided by ETS (1980), validities of .52 and .58 for the GPA alone and the GPA plus the SAT scores respectively, the SAT makes a very modest addition to the prediction of only seven percentage points; the GPA alone explains 27% of the variance in first year college grades against 33.6% explained by the GPA plus the SAT scores. Using the Taylor and Russel (1939) tables, the contribution of the SAT beyond the prediction made by high school grades is only about one or two percentage points; this is the

analysis we need to make in order to determine the goodness of the SAT prediction as compared to the GPA scores alone; if we compare the SAT with chance as Kaplan (1982) did, the SAT is going to come out very well as the GPA scores do.

Predictor vs. Predicted Scores

The selection and admission processes are aimed at reducing the risk of applicant's failure after acceptance. Performance on the criterion instead of performance on the predictor should be the basis of a decision; the preference, however, is toward the predictors because they offer more attractive psychometric characteristics. Minority groups perform lower on the predictor than on the criterion, relatively to the majority group; Gael & Associates (1972, 1974a, 1975b) and Golman and Widawski (1976) report differences between the majority and the minority groups on the predictor larger, sometimes twice as larger, than on the criterion.

Individual vs. Group Fairness

Thorndike (1971) demonstrated convincingly that the predictor may be fair for individuals but not for groups because basing the selection on the predictor, when the validity is less than one, would reject minority candidates who would perform as well or better than some majority candidates if accepted. This realization was the beginning of the short lived group parity models. The models have been rejected, some researchers would say discredited, for being internally inconsistent from a psychometric point of view (Petersen & Novick, 1976) and because they propose different standards for different people (Tenopir, 1981). The models were a solution to a problem, have been rejected on pure theoretical grounds, but no alternative solution has been proposed to deal with the problem the models tried to solve.

A group parity approach to fairness in educational selection has existed in the Nation for many years. In the National Merit Scholarship competition, the number of semifinalists is allocated in proportion to a state's percentage of the Nation's

total graduating class of high school seniors. In a case it is called quotas and in the other state's characteristics (Haney, 1982).

Functional vs. Grade Level Testing

Compensatory educational programs were designed to improve the educational skills, especially reading and math, of students scoring well below their own grade level. Evaluating these programs presented both an educational and a psychometric problem. If students were tested with tests developed for their own grade level, the tasks were above their achievement level creating possible frustration and also questionable test scores because of the large amount of guessing involved when students do not know the answers. If the students were tested at their own functioning level, it was reasoned, both problems would be solved and the functioning level testing was instituted in many schools.

Long, Schaffran, & Kellogg (1977), Flake & Hoover (1979), and Wick (1983) reported that changing from grade to functioning level (a) lowered the average test scores (b) substantially reduced the percentage of chance level scores and (c) they concluded that the functioning level produced more valid scores than the grade level testing. The decision and recommendations are based only on statistical considerations because validity is something more important than just mere reduction of chance errors.

Gonzalez-Tamayo (1983) reported mean group differences of .35GE in Word Knowledge and 1.00GE in Reading between grade and functioning level for the same group of students in fifth grade with only three days between testings. The students scored higher at their own level (pre-test) even though the statistical effect of regression toward the mean suggested the opposite. In another group of second graders, those tested at the functioning level improved only one month in 4.5 months of schooling while those tested at their own grade level improved 0.91GE in Word Knowledge and 0.62GE in Reading. Taking into consideration (1) the magnitude of the difference (2) the fact that tests of achievement include what is taught in class

and (3) the average amount of correct responses (about 40%) was well above chance (25%) and also above the rule of thumb suggested by some researchers (33%), it was concluded that grade level testing is a better measure of the students than functioning level.

All reports indicate that students in compensatory educational programs do better when tested at their own grade level than when they are tested at the functioning level. Groups always score higher on tests developed for their own group than on tests developed for other groups (Anastasi, 1982). If the elimination of chance scores were our only objective in a test, we could easily achieve it by testing students with lower levels but the group mean would also be lower. A national sample of ninth graders taking the Elementary Level of the MAT (1971) have to miss, as an average, not more than one or two items in order to get the same mean score they would get if tested by the Intermediate Level (grade level), a very unlikely event. As a result, it seems reasonable to say that any group, not only those in compensatory education, would score lower in a test outside their own grade than on a test at their grade level.

CONCLUSION

The controversy about aptitude testing started with questions raised about test fairness and bias because of their adverse impact on minorities. One of the reasons why we are not closer to a solution could be that beliefs, at least in a subtle way, have influenced the arguments and beliefs are not easily changed by the mere force of reason. Tests were developed by the majority and for the majority and there are serious doubts about their fairness and even usefulness when they are used with minorities. Psychological measurements are not and never have been precise techniques as we would like them to be or as some of our statements make them appear to be.

In the seventies, a drastic decline in the SAT scores was discovered and the report of the panel selected to study the problem mentioned as possible causes for the decline, the composition of the population taking the test (more and less prepared individuals), the schools, the family, the students themselves, and even the country (Wirtz, 1977). It was never suggested that the possible cause could be that the content of the test was less relevant to today's schooling than it was few years ago. Today schools teach different subjects and emphasize different areas; society and the communication media do not show interest in the same things or in the same degree than they did years ago. Today students may have the same ability but in different areas. Could it be possible that the SAT is, in some way, out of touch with society and the new reality? This is only a possibility like the causes mentioned in the report; nevertheless, anything and anybody was blamed for the decline except the test itself. Flynn (1984), on the other hand, reported massive mean IQ gains from 1932 to 1978 of 13.8 points or about .86 standard deviations. This increase, if real, indicates a tremendous boost of the intelligence in the population as we have not seen in a developed society and reveals the inexplicable combination of IQ gains and SAT scores losses.

During the whole controversy, the reaction has been to stick to the tests and find some other explanations for the problems. The scientific community should be more receptive and less defensive of criticism about ability testing and show some sensitivity toward the problems created by our imperfect instruments. The responsibility for new ideas is for every psychologist but it is more pressing for those who benefit the most from the tests: test industry and large companies. These two groups are very often the only ones with the data and monetary resources necessary to carry out a serious study of all the issues raised during the controversy. Data come to us in response to a question; the tragedy is that, many times, those who have the data do not ask the proper questions and those who ask the proper questions can not get the data.

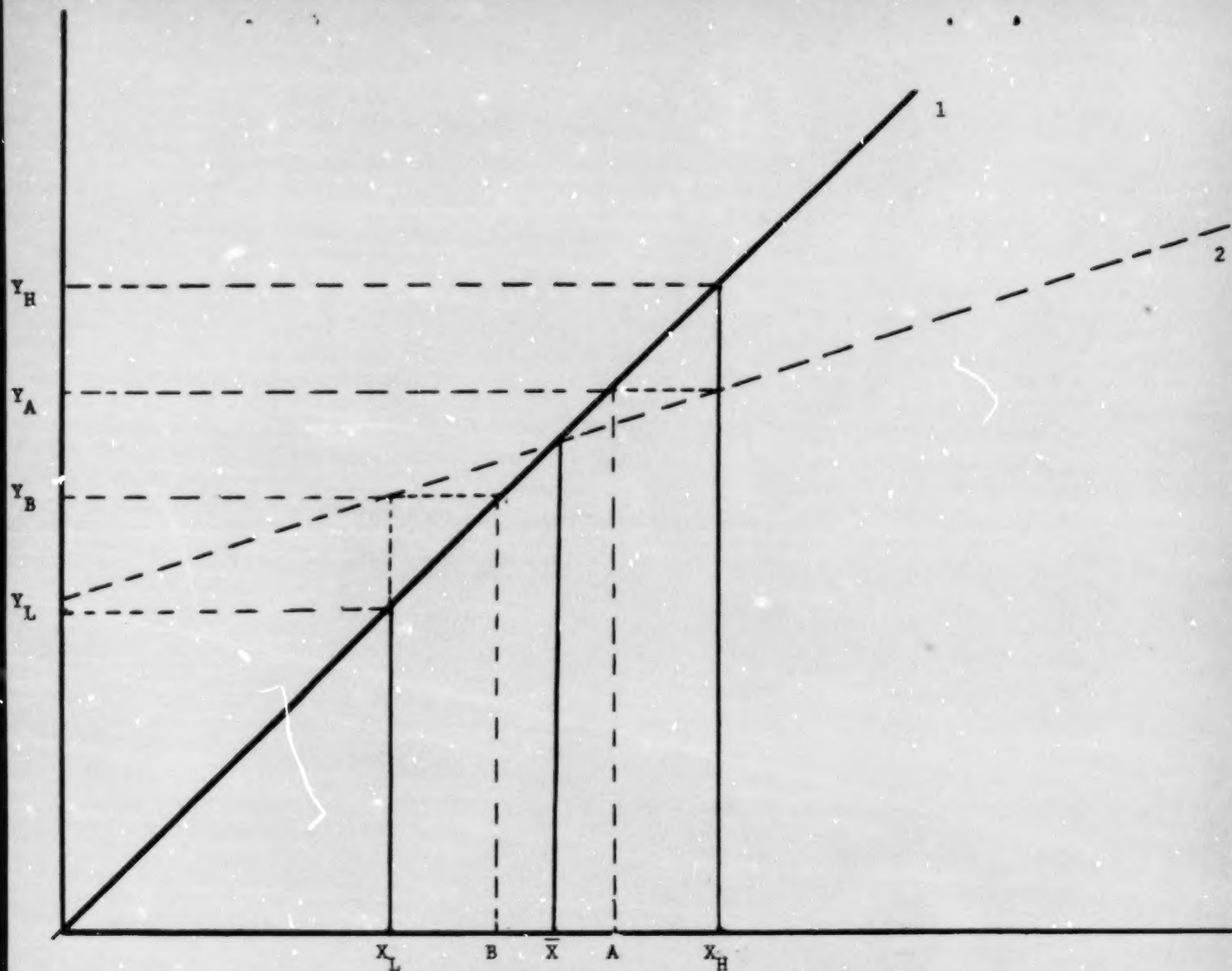


Figure 1: Graphic Representation of the Prediction when the Reliability or Validity are One (1) and when they are Less than Perfect (2)

BIBLIOGRAPHY

- Anastasi, A. Psychological Testing (5th Ed.). McMillan Publishing Co., Inc. New York, 1982.
- Berry, J.W. On Cross-cultural Comparability. International Journal of Psychology, 1969, 4, 119-128.
- Board of Education. Language Assesment Battery. New York City, 1976.
- Behm, A.E. Test of Basic Concepts. The Psychological Corporation, N.Y., 1971.
- Breland, H.M. Population Validity and College Entrance Measures. The College Board, New York, 1979.
- Carlson, J.S. Children's Probability Judgement as Related to Age, Socioeconomic Level, and Sex. Human Development, 1969, 12, 192-203.
- Cleary, T.A. Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, 1968, 5, 115-124.
- Cleary, T.A., Humphreys, L., Wesman, A.G., & Kendrick, S.A. Educational Uses of Tests with Disadvantaged Students. American Psychologist, 1975, 30, 15-41.
- Cole, M., Hood, L., & McDermott, R. Ecological Niche Picking: Ecological Invalidity as an Axiom of Experimental Cognitive Psychology. Laboratory of Comparative Human Cognition, Rockefeller University, New York, 1978.
- Cronbach, L.J. Validation. In Thorndike, R.L. (Ed.) Educational Measurement. Washington D.C.: American Council on Education, 1971.
- Educational Testing Service. Test Use and Validity. ETS, Princeton, N.J., 1980.
- Farr, J.L., O'Leary, B.S., Pfeiffer, C.M., Dolstein, I.L., & Bartlett, C.J. Ethnic Group Membership as a Moderator in the Prediction of Job Performance: An Examination of Some Less Traditional Predictors. AIR Technical Report No.2. Washington D.C.: American Institute for Research, September 1971.
- Flaugher, R.L. The many Definitions of Test Bias. American Psychologist, 1978, 33, 671-679.

- Flynn, J.R. The Mean IQ of Americans: Massive Gains 1932 to 1978. Psychological Bulletin, 1984, 95, 29-51.
- Gael, S. & Grant, D.L. Employment Test Validation for Minority and Non-minority Telephone Company Service Representatives. Journal of Applied Psychology, 1972, 56, 135-139.
- Gael, S., Grant, D.L., & Ritchie, R.J. Employment Test Validation for Minority and Non-minority Telephone Company Operators. Journal of Applied Psychology, 1975, 60, 411-419 (a).
- Gael, S., Grant, D.L., & Ritchie, R.J. Employment Test Validation for Minority and Non-minority Clerks with Work Sample Criteria. Journal of Applied Psychology, 1975, 60, 420-426 (b).
- Giselli, E. The Validity of Occupational Aptitude Tests. New York: Wiley, 1966.
- Giselli, E. The Validity of Aptitude Tests in Personnel Selection. Personnel Psychology, 1973, 26, 461-477.
- Goodenough, F.L. & Harris, D.B. Studies in the Psychology of Children's Drawings: 1928-1949. Psychological Bulletin, 1950, 47, 369-433.
- Goldman, R.D. & Slaughter, R.E. Why College Grade Point Average is Difficult to Predict. Journal of Educational Psychology, 1976, 68, 9-14.
- Goldman, R.D. & Widawski, M.H. An Analysis of Types of Errors in the Selection of Minority College Students. Journal of Educational Measurement, 1976, 13, 186-200.
- Doldschmidt, W. Comparative Functionalism. Berkeley, University of California Press, 1966.
- Gonzalez-Tamayo, E. Bilingual Tests: a Joke or an Insult? Internal Research Report, PRC Metronomics, Inc., New York, 1980.
- Gonzalez-Tamayo, E. Functioning-Level and Grade-Level Testing in Compensatory Educational Programs. Internal Research Report, PRC Metronomics, Inc. N.Y., 1983.
- Green, B.F., Jr. In Defense of Measurement. American Psychologist, 1978, 33, 664-670.
- Haney, W.M. Disabilities of Committee Testing: a Review. Educational Measurement, Issues and Practices, 1982, 1, 13-18.

- Hoffman, B. The Tyranny of Testing. New York: Crowell-Collier, 1962.
- Hogan, R., DeSoto, C.B., & Solano, C. Traits, Tests, and Personality Research. American Psychologist, 1977, 32, 255-264.
- Holtzman, W.P. Cross Cultural Studies in Psychology. International Journal of Psychology, 1968, 3, 83-91.
- Hunter, J.E. & Schmidt, F.L. A Critical Analysis of the Statistical and Ethical Implications of Five Definitions of Test Fairness. Psychological Bulletin, 1976, 83, 1053-1071.
- Irvine, S.H. Factor Analysis of African Abilities and Attainments: Constructs across Cultures. Psychological Bulletin, 1969, 71, 20-32.
- Israel, J. & Rommetveit, R. Notes on the Standardization of Experimental Manipulations and Measurement in Cross-national Research. Journal of Social Issues, 1954, 10, 61-68.
- Jensen, A.R. Bias in Mental Testing. The Free Press, New York, 1980.
- Kaplan, A. The Conduct of Inquiry. San Francisco: Chandler, 1964.
- Kaplan, R.M. Nader's Raid on the Testing Industry. American Psychologist, 1982, 37, 15-23.
- Lerner, B. The War on Testing: Detroit Edison in Perspective. Princeton, New Jersey, ETS, 1973.
- Levine, M.V. & Drasgow, F. The Relation between Incorrect Option Choice and Estimated Ability. Educational and Psychological Measurement, 1983, 43, 675-685.
- Linn, R.L. Admission Testing on Trial. American Psychologist, 1982, 37, 279-291.
- Linn, R.L. Single-group Validity, Differential Validity, and Differential Prediction. Journal of Applied Psychology, 1978, 63, 507-512.
- Long, J.V., Schaffran, J.A. & Kellogg, T.M. Effects of Out-of-level Survey Testing on Reading Achievement Scores of Title I, ESEA Students. Journal of Educational Measurement, 1977, 14, 203-214.
- McNemar, Q. The Revision of the Stanford-Binet Scale. Boston: Houghton-Mifflin, 1942.
- McClellan, D.C. Testing for Competence rather than for Intelligence. American Psychologist, 1973, 28, 1-14.

- Messick, S. Test Validity and the Ethics of Assessment. American Psychologist, 1980, 25, 1012-1027.
- Munday, L.A. Factors Influencing the Predictability of College Grades. American Educational Research Journal, 1970, 7, 99-107.
- Metropolitan Achievement Tests (MAT). New York: Harcourt Brace Jovanovich, 1971
- Nisbert, R. Sociology As an Art Form. New York: Oxford University Press, 1976.
- Novick, M.R., & Ellis, D.D., Jr. Equal Opportunity in Educational and Employment Selection. American Psychologist, 1977, 32, 306-320.
- Olmedo, E.L. Testing Linguistic Minorities. American Psychologist, 1981, 36, 1078-1085.
- Petersen, N.S. & Novick, M.R. An Evaluation of Some Models for Culture-fair Selection. Journal of Educational Measurement, 1976, 13, 3-29.
- Plake, B.S. & Hoover, H.D. The Comparability of Equal Raw Scores Obtained from It-level and Out-of-level Testing: One Source of the Discrepancy between In-level and Out-of-level Grade Equivalent Scores. Journal of Educational Measurement, 1979, 16, 271-278.
- Printice Hall and American Society for Personnel Administration. Survey Probes Employee Testing and Selection Procedures, Personnel Management Policies and Practices Report. Englewood Cliffs, N.J.: Printice Hall, 1975.
- Samuda, R.J. Psychological Testing of American Minorities: Issues and Consequences. New York: Dodd, Mead, and Co., 1975.
- Siegel, A.I. The Miniature Job Training and Evaluation Approach: Additional Findings. Personnel Psychology, 1983, 36, 41-56.
- Schmidt, F.L., Hunter, J.E., McKenzie, R.C., & Muldrow, T.W. Impact of Valid Selection Procedures on Work-force Productivity. Journal of Applied Psychology, 1979, 64, 609-625.
- Schmidt, F.L., Gas Rosenberg, I., & Hunter, J.E. Validity Generalization Results for Computer Programmers. Journal of Applied Psychology, 1980, 65, 643-661.
- Schmidt, F.L., & Hunter, J.E. Employment Testing: Old Theories and New Findings. American Psychologist, 1981, 36, 1128-1137 (a).

- Schmidt, F.L., Hunter, J.E., & Caplan, J.R. Validity Generalization Results for Two Groups in the Petroleum Industry. Journal of Applied Psychology, 1981, 66, 261-273 (b).
- Schmidt, F.L., Hunter, J.E., & Pearlman, K. Task Differences as Moderator of Aptitude Test Validity in Selection: a Red Herring. Journal of Applied Psychology, 1981, 66, 166-185 (c).
- Sherwood, J.J., & Natauski, M. Predicting Conclusions of Negro-White Intelligence Research from Biographical Characteristics of the Investigator. Journal of Personality and Social Psychology, 1968, 8, 53-58.
- Stanley, J.C. & Porter, A.C. Correlation of SAT Scores with College Grades for Negroes versus Whites. Journal of Educational Measurement, 1967, 4, 199-218.
- Taylor, H.C., & Russell, J.T. The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables. Journal of Applied Psychology, 1939, 565-578.
- Tenopyr, M.L. The Realities on Employment Testing. American Psychologist, 1981, 36, 1120-1127.
- Thorndike, R.L. Concepts of Culture-Fairness. Journal of Educational Measurement, 1971, 8, 63-70.
- Vaudenberg, S.G. The Primary Mental Abilities of South-American Students: a Second Comparative Study of the Generality of a Cognitive Factor Structure. Multivariate Behavioral Research, 1967, 2, 175-198.
- Weitzman, R.A. The Prediction of College Achievement by the Scholastic Aptitude Test and the High School Record. Journal of Educational Measurement, 1982, 19, 179-191.
- Wigdor, A.K., & Garner, W.R. (Eds.). Ability Tests: Uses, Consequences, Controversies. Washington D.C.: National Academy Press, 1982.
- Wick, J.W. Reducing Proportion of Chance Scores in Inner-City Standardized Testing Results: Impact on Average Scores. American Educational Journal, 1983, 20, 461-463.
- Wirtz, W. On Further Examination: Report of the Panel on Scholastic Aptitude Test Scores Decline. Princeton, N.J.: College Entrance Examination Board, 1977.
- Wolf, R.M. Achievement in America: National Report of the U.S. for the International Educational Achievement Project. New York: Teachers College Press, 1977.